DDS Analkytics

Daniel Davieau, LAkeitra Webb, Emil Ramos

April 18, 2018

Info

rm(list=ls())  
setwd("~/GitHub/DDS-Analytics")  
library(knitr)  
library(readxl)  
library(dygraphs)  
library(tidyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(forecast)  
library(xtable)  
library(kableExtra)  
library(readr)  
library(reshape2)

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths

library(wordcloud2)  
#sessionInfo()  
palette <- c("#999999", "#E69F00", "#009E73", "#0072B2", "#D55E00", "#CC79A7")  
#Orange "#E69F00"  
#Green "#009E73"  
#Blue "#0072B2"  
#DrkOran"#D55E00"  
#Pink "#CC79A7"

Oservations and Variables in the data that was provided:

CaseStudy2\_data <- read\_excel("Input/CaseStudy2-data.xlsx")  
dims<-as.data.frame(dim(CaseStudy2\_data))  
cbind.data.frame("Observations",dims)

## "Observations" dim(CaseStudy2\_data)  
## 1 Observations 1470  
## 2 Observations 35

rownames(dims)<-c("Observations","Variables")  
colnames(dims)<-c("Count")  
kable(dims, "html") %>%  
kable\_styling(bootstrap\_options = "striped", full\_width = F)

Count

Observations

1470

Variables

35

Preliminary data analysis; Data Structure, Quality Analysis etc. b. The column names are either too much or not enough. Change the column names so that they do not have spaces, underscores, slashes, and the like. All column names should be under 12 characters. Make sure you’re updating your codebook with information on the tidied data set as well.

class((CaseStudy2\_data))

## [1] "tbl\_df" "tbl" "data.frame"

unique(CaseStudy2\_data$Gender)

## [1] "Female" "Male"

unique(CaseStudy2\_data$YearsAtCompany)

## [1] 6 10 0 8 2 7 1 9 5 4 25 3 12 14 22 15 27 21 17 11 13 37 16  
## [24] 20 40 24 33 19 36 18 29 31 32 34 26 30 23

unique(CaseStudy2\_data$Attrition)

## [1] "Yes" "No"

a Remove all observations where the participant is under age 18. No further analysis of underage individuals is permitted by your client. Remove any other age outliers as you see fit, but be sure to tell what you’re doing and why.

#verify no records with participants under age 18  
underage<-as.matrix(min(CaseStudy2\_data$Age))  
colnames(underage)<-c("Youngest Age of Participants")  
kable(underage, "html") %>%  
kable\_styling(bootstrap\_options = "striped", full\_width = F)

Youngest Age of Participants

18

b Please provide (in pretty-fied table format or similar), descriptive statistics on at least 7 variables (age, Income, etc.).

# Interesting variables: daily rate, education, job satisfaction,number of companies worked, performance rating,years at company, years with current manager  
# Variables to avoid:Age, Gender, marital status, relationship statisfaction, total working years  
  
#age<-matrix(summary(CaseStudy2\_data$Age))  
dr <- matrix(summary(CaseStudy2\_data$DailyRate))  
cw<-matrix(summary(CaseStudy2\_data$NumCompaniesWorked))  
yac<-matrix(summary(CaseStudy2\_data$YearsAtCompany))  
ywm<-matrix(summary(CaseStudy2\_data$YearsWithCurrManager))  
dfh<-matrix(summary(CaseStudy2\_data$DistanceFromHome))  
psh<-matrix(summary(CaseStudy2\_data$PercentSalaryHike))  
ycr<-matrix(summary(CaseStudy2\_data$YearsInCurrentRole))  
  
x<-data.frame(cbind(dr,cw,yac,ywm,dfh,psh,ycr))  
colnames(x)<-c("DailyRate","Number of Companies Worked","Years at Company","Years with Manager", "Distance From Home","Percent Salary Hike","Years in Current Role")  
rownames(x)<-c("Min.","1st Qu.","Median","Mean","3rd Qu.","Max.")  
kable(x, "html") %>%  
kable\_styling(bootstrap\_options = "striped", full\_width = F)

DailyRate

Number of Companies Worked

Years at Company

Years with Manager

Distance From Home

Percent Salary Hike

Years in Current Role

Min.

102.0000

0.000000

0.000000

0.000000

1.000000

11.00000

0.000000

1st Qu.

465.0000

1.000000

3.000000

2.000000

2.000000

12.00000

2.000000

Median

802.0000

2.000000

5.000000

3.000000

7.000000

14.00000

3.000000

Mean

802.4857

2.693197

7.008163

4.123129

9.192517

15.20952

4.229252

3rd Qu.

1157.0000

4.000000

9.000000

7.000000

14.000000

18.00000

7.000000

Max.

1499.0000

9.000000

40.000000

17.000000

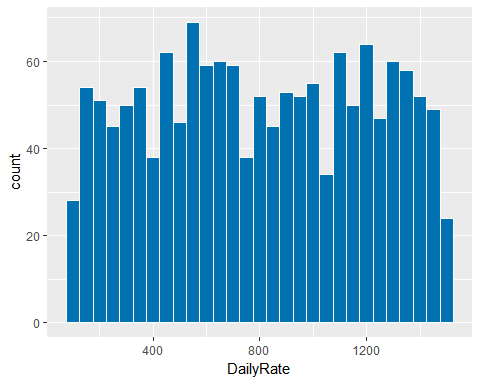
29.000000

25.00000

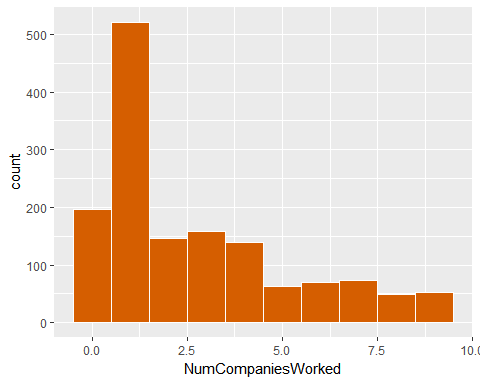
18.000000

Create a simple histogram for two of them. Comment on the shape of the distribution in your markdown.

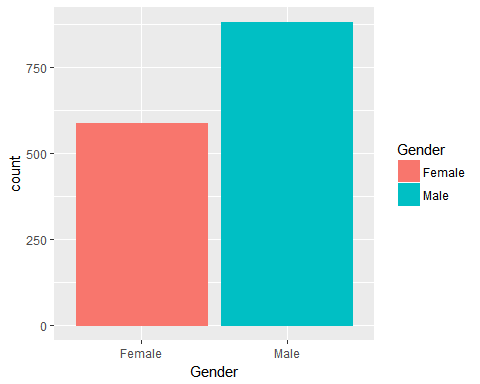
ggplot(CaseStudy2\_data, aes(x=DailyRate, color=palette))+  
geom\_histogram(binwidth=50, color="white" , fill="#0072B2")



ggplot(CaseStudy2\_data, aes(x=NumCompaniesWorked))+  
 geom\_histogram( binwidth=1, position="identity",color="white", fill="#D55E00")

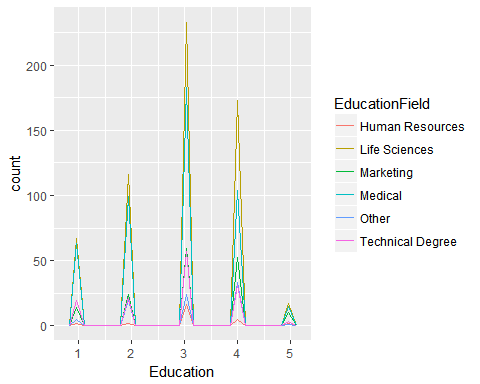
 c Give the frequencies (in table format or similar) for Gender, Education, and Occupation. They can be separate tables, if that’s your choice.

ggplot(CaseStudy2\_data, aes(x=Gender, color = Gender, fill=Gender)) +geom\_bar(stat = "count")

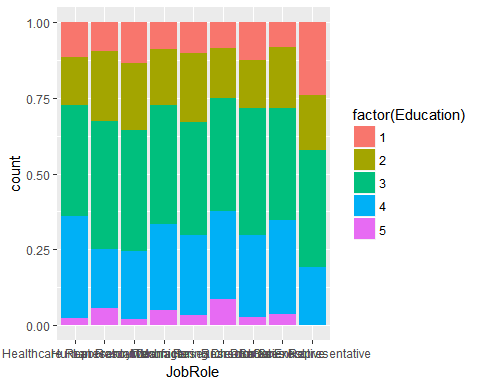


#Orange "#E69F00"  
#Green "#009E73"  
#Blue "#0072B2"  
#DrkOran"#D55E00"  
#Pink "#CC79A7"  
  
ggplot(CaseStudy2\_data, aes(Education, colour = EducationField)) +  
 geom\_freqpoly(show.legend = TRUE, inherit.aes = TRUE)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# ggplot(CaseStudy2\_data, aes(x=JobRole, colour = Gender)) +  
# geom\_freqpoly(stat = "count", show.legend = TRUE, inherit.aes = TRUE)  
  
ggplot(CaseStudy2\_data) +  
 aes(x = JobRole, fill = factor(Education)) +  
 geom\_bar(position = "fill")



head(CaseStudy2\_data)

## # A tibble: 6 x 35  
## Age Attrition BusinessTravel DailyRate Department DistanceFromHome  
## <dbl> <chr> <chr> <dbl> <chr> <dbl>  
## 1 41. Yes Travel\_Rarely 1102. Sales 1.  
## 2 49. No Travel\_Frequen~ 279. Research & D~ 8.  
## 3 37. Yes Travel\_Rarely 1373. Research & D~ 2.  
## 4 33. No Travel\_Frequen~ 1392. Research & D~ 3.  
## 5 27. No Travel\_Rarely 591. Research & D~ 2.  
## 6 32. No Travel\_Frequen~ 1005. Research & D~ 2.  
## # ... with 29 more variables: Education <dbl>, EducationField <chr>,  
## # EmployeeCount <dbl>, EmployeeNumber <dbl>,  
## # EnvironmentSatisfaction <dbl>, Gender <chr>, HourlyRate <dbl>,  
## # JobInvolvement <dbl>, JobLevel <dbl>, JobRole <chr>,  
## # JobSatisfaction <dbl>, MaritalStatus <chr>, MonthlyIncome <dbl>,  
## # MonthlyRate <dbl>, NumCompaniesWorked <dbl>, Over18 <chr>,  
## # OverTime <chr>, PercentSalaryHike <dbl>, PerformanceRating <dbl>,  
## # RelationshipSatisfaction <dbl>, StandardHours <dbl>,  
## # StockOptionLevel <dbl>, TotalWorkingYears <dbl>,  
## # TrainingTimesLastYear <dbl>, WorkLifeBalance <dbl>,  
## # YearsAtCompany <dbl>, YearsInCurrentRole <dbl>,  
## # YearsSinceLastPromotion <dbl>, YearsWithCurrManager <dbl>

d Give the counts (again, pretty table) of management positions.

c Give the frequencies (in table format or similar) for Gender, Education, and Occupation. They can be separate tables, if that’s your choice. d Give the counts (again, pretty table) of management positions.

a Note: You should make all of these appealing looking. Remember to include things like a clean, informative title, axis labels that are in plain English, and readable axis values that do not overlap. b Create barcharts in ggplot or similar The bars should be in descending order, Use any color palette of your choice other than the default. c Is there a relationship between Age and Income? Create a scatterplot and make an assessment of whether there is a relationship. Color each point based on the Gender of the participant. You’re welcome to use lm() or similar functions to back up your claims.

d What about Life Satisfaction? Create another scatterplot. Is there a discernible relationship there to what?

The executive leadership has identified predicting employee turnover as its first application of data science for talent management. Before the business green lights the project, they have tasked your data science team to conduct an analysis of existing employee data.

determine factors that lead to attrition identify (at least) the top three factors that contribute to turnover The business is also interested in learning about any job role specific trends that may exist in the data set (e.g., “Data Scientists have the highest job satisfaction”

• Title • Authors all listed • Presentation Outline • Business Objectives • Data Sourced • Methodology • Evaluation/Results • Summary • Business Objectives • Data Source • Where you got it • Basic Statistics • Methodology • Steps • Workflow • Evaluation/Results • Tell me the percentages • Show me graphs with explanations • The top three factors that contribute to turnover • Tell me about any job role specific trends that may exist in the data set • Provide any other interesting trends and observations from your analysis • Other things to consider? • Summary • Insights • Recommendations • Improvements • Questions Have your presentation as a PDF ready to present on April 24.

c Some columns are, due to Qualtrics, malfunctioning.  
d Make sure your columns are the proper data types (i.e., numeric, character, etc.). If they are incorrect, convert them.